An algorithm for optimal fusion of atlases with different labeling protocols

Eugenio Iglesias, Juan; Sabuncu, Mert Rory; Aganj, Iman; Bhatt, Priyanka; Casillas, Christen; Salat, David; Boxer, Adam; Fischl, Bruce; Van Leemput, Koen

Published in:
NeuroImage

Link to article, DOI:
10.1016/j.neuroimage.2014.11.031

Publication date:
2015

Document Version
Publisher's PDF, also known as Version of record

NeuroImage 106 (2015) 451–463

Contents lists available at ScienceDirect

NeuroImage

journal homepage: www.elsevier.com/locate/ynimg

An algorithm for optimal fusion of atlases with different labeling protocols

Juan Eugenio Iglesias [a,*], Mert Rory Sabuncu [b,d], Iman Aganj [b], Priyanka Bhatt [c], Christen Casillas [c], David Salat [b], Adam Boxer [c], Bruce Fischl [d,b], Koen Van Leemput [b,e,f,g]

[a] Basque Center on Cognition, Brain and Language (BCBL), Spain
[b] Athinoula A. Martinos Center for Biomedical Imaging, Harvard Medical School/Massachusetts General Hospital, Charlestown, MA, USA
[c] Memory and Aging Center, University of California, San Francisco, USA
[d] MIT Computer Science and Artificial Intelligence Laboratory (CSAIL), USA
[e] Department of Applied Mathematics and Computer Science, Technical University of Denmark, Denmark
[f] Department of Information and Computer Science, Aalto University, Finland
[g] Department of Biomedical Engineering and Computational Science, Aalto University, Finland

ARTICLE INFO

Article history:
Accepted 14 November 2014
Available online 22 November 2014

Keywords:
Segmentation
Label fusion

ABSTRACT

In this paper we present a novel label fusion algorithm suited for scenarios in which different manual delineation protocols with potentially disparate structures have been used to annotate the training scans (hereafter referred to as "atlases"). Such scenarios arise when atlases have missing structures, when they have been labeled with different levels of detail, or when they have been taken from different heterogeneous databases. The proposed algorithm can be used to automatically label a novel scan with any of the protocols from the training data. Further, it enables us to generate new labels that are not present in any delineation protocol by defining intersections on the underling labels. We first use probabilistic models of label fusion to generalize three popular label fusion techniques to the multi-protocol setting: majority voting, semi-locally weighted voting and STAPLE. Then, we identify some shortcomings of the generalized methods, namely the inability to produce meaningful posterior probabilities for the different labels (majority voting, semi-locally weighted voting) and to exploit the similarities between the atlases (all three methods). Finally, we propose a novel generative label fusion model that can overcome these drawbacks. We use the proposed method to combine four brain MRI datasets labeled with different protocols (with a total of 102 unique labeled structures) to produce segmentations of 148 brain regions. Using cross-validation, we show that the proposed algorithm outperforms the generalizations of majority voting, semi-locally weighted voting and STAPLE (mean Dice score 83%, vs. 77%, 80% and 79%, respectively). We also evaluated the proposed algorithm in an aging study, successfully reproducing some well-known results in cortical and subcortical structures.

Introduction

Automatic segmentation of brain structures from MRI data makes it possible to carry out neuroimaging studies at larger scales than manual tracings would, since the latter are very time consuming to make. Moreover, automatic segmentation methods are also more repeatable and reliable than their manual counterparts. Brain MRI segmentation has been used in a number of applications, such as tractography (Yendiki et al., 2011), surgical planning (Cline et al., 1990) and studies of aging (Walhovd et al., 2005), brain development (Knickmeyer et al., 2008) and pathologies like Alzheimer's disease (De Jong et al., 2008).

One family of supervised segmentation techniques that has become popular in brain MRI is multi-atlas segmentation (Rohlfing et al., 2004). In conventional atlas-based segmentation, the grayscale image of the atlas is nonlinearly registered to the space of the test scan, and the resulting transform is then used to warp the corresponding labels, which provide an estimate of the segmentation. Since a single atlas is not sufficient to cover the whole spectrum of variability within a population, multi-atlas segmentation has emerged as a natural extension. Using multiple atlases, this family of techniques produces more accurate segmentations (Awate and Whitaker, 2014) by: (1) independently registering several atlases to the test scan; (2) using the resulting transforms to deform the corresponding label images; and (3) combining the registered label maps into a single estimate of the segmentation with a label fusion algorithm. Multi-atlas segmentation is becoming widespread for three reasons. First, the maturity of registration algorithms (e.g., ANTs/SyN (Avants et al., 2008), Elastix (Klein et al., 2010)) enables multi-atlas techniques to achieve very high performance. Second, the public availability of such methods makes multi-atlas segmentation

* Corresponding author.
E-mail address: e.iglesias@bcbl.eu (J.E. Iglesias).

http://dx.doi.org/10.1016/j.neuroimage.2014.11.031

easy to implement. And third, the relative computational cost associated with nonlinearly registering the atlases is quickly diminishing thanks to the rapid increase in processing power of computers.

The choice of label fusion method is critical for the performance of multi-atlas segmentation. Early algorithms include *best atlas selection* (Rohlfing et al., 2004) and *majority voting* (Heckemann et al., 2006). The former estimates the segmentation as the labels of the atlas that is most similar to the test scan after registration. In this context, similarity can be measured with the same metrics that are typically used in image registration, such as cross-correlation, mutual information or sum of squared differences. Majority voting, on the other hand, operates at the voxel level by locally assigning the most frequent deformed atlas label at each spatial location – without considering the image intensity information. The performance of majority voting can be increased by an atlas selection process, in which only the deformed atlases that are most similar to the target scan are considered in the fusion (Aljabar et al., 2009; Duc et al., 2013).

Later fusion methods compute the segmentation as a weighted combination of the labels of the registered atlases such that higher weights are given to more similar atlases. The weights can be global (Artaechevarria et al., 2008) or local (Isgum et al., 2009; Coupé et al., 2011; Wang et al., 2013; Sabuncu et al., 2010). *Sabuncu* et al. (Sabuncu et al., 2010) have shown that many of these multi-atlas methods can be written within a unified generative model. Another popular label fusion approach is STAPLE (Warfield et al., 2004) and its variants (Asman and Landman, 2012, 2013; Cardoso et al., 2013; Akhondi-Asl and Warfield, 2013); while this method was originally developed to combine multiple manual segmentations from different human raters, it is increasingly being applied in the context of multi-atlas label fusion.

All the aforementioned label fusion algorithms assume that all structures of interest are labeled in all atlases, which is a rather limiting constraint. Eliminating this requirement would have several practical implications:

- It would enable us to combine training scans from different datasets even if they have different sets of annotated structures. In turn, this would make it possible to take advantage of the increasing amount of heterogeneously labeled MRI data that are publicly available.
- It would also enable us to segment structures that are not included in any of the datasets, but defined as the intersection of labels. For instance, the intersection of the lateral postcentral region and the cerebral gray matter would define the primary somatosensory cortex.
- It would allow for the fusion of segmentations from different modalities with different field of views and resolution. For instance, it would be possible to combine standard resolution brain MRI (1 mm resolution) with high-resolution MRI with limited field of view or even histology or optical coherence tomography data.
- It would be useful if one were to manually relabel a subset of atlases to include finer structures in the annotations. For example, in a large dataset with the hippocampi already labeled, an expert rater can additionally delineate the hippocampal subfields – which is extremely difficult and time consuming – in just a few cases. Traditional label fusion methods would only be able to use these few scans in the segmentation, having to disregard the information in all the scans in which the subfields are not labeled.

Despite the practical implications that a label fusion algorithm which allows for heterogeneously labeled atlases would have, this direction remains largely unexplored in the literature. To the best of our knowledge, only a particular case of label fusion with heterogeneous labels has been considered so far: the situation in which some of the labels are missing in some of the atlases. To tackle this problem, *Landman* et al. (Landman et al., 2009, 2010, 2012) propose an ad-hoc solution by modifying the STAPLE framework such that unlabeled voxels are ignored and the confusion matrix entries corresponding to the missing structures are fixed. *Commowick* et al. (Commowick et al., 2012) propose ameliorating the effect of missing labels by adding a prior on the confusion matrices to the STAPLE algorithm that, when a label is missing, encourages higher a transition probability from that label to the background. However, such an approach treats as background all the voxels that have not been labeled with one of the foreground labels.

In this study, we present a family of probabilistic models for label fusion that make it possible to use atlases that have been annotated with different protocols. In our models, the atlases are assumed to have a hidden "fine" segmentation with all the structures present in the training data – including those defined by intersections of labels. The actual observed labels are assumed to have been obtained by collapsing groups of hidden fine labels into more general, coarse labels.

The contribution of this study is twofold:

i. We use probabilistic models of label fusion to extend three popular methods (majority voting, semi-locally weighted fusion and STAPLE) to the scenario of heterogeneously labeled atlases.
ii. We propose a new generative model for label fusion that can overcome the limitations of these generalizations – the inability to produce meaningful posteriors and to exploit the similarities between the atlases – and show that it outperforms the generalizations in experiments with four datasets.

The rest of this paper is organized as follows. In the Methods section, we describe the general framework for label fusion with heterogeneously labeled atlases, propose the generalizations of the different methods, identify their disadvantages, and present a new fusion algorithm to address their shortcomings. In the Experiments and results section, we assess the performance of the different algorithms with experiments on four different datasets. Finally, the Conclusion and discussion section closes the paper.

Methods

In this section, we first introduce the general framework and define the variables that we will use throughout the paper. Then, we present the generalizations of majority voting, semi-locally weighted voting and STAPLE and identify their weaknesses. Finally, we introduce a label fusion method that addresses these shortcomings.

General framework

Throughout the remainder of this paper, we will assume that a test scan consisting of J voxels is to be segmented. We will use $\mathbf{y} = \{y_j, j = 1, \ldots, J\}$ to refer to the image intensities, and $\mathbf{s} = \{s_j, j = 1, \ldots, J\}$ to refer to its hidden, underlying segmentation. Let us also assume that a set of N atlases has been pre-registered to a test scan with a non-linear algorithm. Let $\{\mathbf{i}_n\}$ (where $\mathbf{i}_n = \{i_{nj}, j = 1, \ldots, J\}$) be the observed image intensities of the N registered atlases, and let $\{\mathbf{l}_n\}$ (where $\mathbf{l}_n = \{l_{nj}, j = 1, \ldots, J\}$) be the corresponding discrete labels, defined at the finest detail level. Their values range from 1 to L, the total number of fine labels.

These deformed labels $\{\mathbf{l}_n\}$ are not directly observed; instead, we have access to a different set of coarse labels $\{\mathbf{c}_n\}$ (where $\mathbf{c}_n = \{c_{nj}, j = 1, \ldots, J\}$), which correspond to the actual manual delineations. The coarse labels $\{\mathbf{c}_n\}$ are obtained by collapsing the fine labels $\{\mathbf{l}_n\}$ into different groups of labels by means of a set of N deterministic, protocol-specific functions: $c_{nj} = f_n(l_{nj})$. A protocol function could, for instance, collapse the hippocampal subfields into a single hippocampal label. Having a separate f_n for each atlas enables us to combine different labeling protocols. Different protocol functions can collapse the same fine label into different coarse labels; for instance, orbital cortex could be collapsed into the cerebral cortex by one protocol and into the frontal lobe by another.

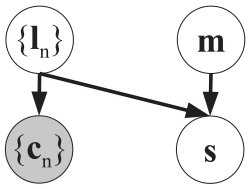

Fig. 1. Graphical model for generalized majority voting. Shaded variables are observed.

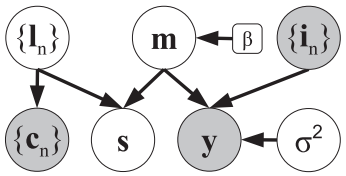

Fig. 2. Graphical model for generalized semi-locally weighted voting. Shaded variables are observed.

We will now generalize three existing models – majority voting, locally weighted voting and STAPLE – to the scenario with collapsed labels. In all cases, the original algorithm (i.e., without heterogeneously labeled atlases) is recovered when all protocol functions are bijective (i.e., no labels are collapsed).

Generalization of majority voting

As explained in Sabuncu et al. (2010), majority voting can be seen as the most likely labeling in a probabilistic model in which the segmentation s_j is sampled randomly from one of the N atlases, as indexed by a hidden discrete field $\mathbf{m} = \{m_j, j = 1, ..., J\}$. Specifically, the value of the field at a certain voxel $m_j \in \{1, ..., N\}$ indexes from which atlas we take the label at voxel j, i.e., $s_j = l_{m_j j}$. The field $\mathbf{m}$ follows a flat prior, i.e., $p(\mathbf{m}) \propto 1$. It is straightforward to show that, after marginalizing over $\mathbf{m}$, the distribution of the segmentation at a given location is equal to its relative frequency within the propagated labels of the atlases. Therefore, the most likely segmentation is the mode of the propagated labels.

The graphical model for generalized majority voting is shown in Fig. 1, and the corresponding equations in Table 1; note that we have assumed a flat prior over the hidden labels. The main difference with the original model is that the fine labels of the atlases are not available; instead, we observe their coarse labels – but still want to compute the segmentation at the fine level. To find the most likely segmentation within this model, we can operate at each spatial location j independently – since the model factorizes over voxels. The problem to solve is $\hat{s}_j = argmax_{s_j} p(s_j|c_{\cdot j})$, where $c_{\cdot j}$ denotes all the coarse atlas labels at voxel j. The probability $p(s_j|c_{\cdot j})$ is given by:

$$\begin{aligned} p\left(s_j|c_{\cdot j}\right) &= \sum_{m_j} p\left(m_j\right) p\left(s_j|m_j, c_{\cdot j}\right) \\ &= \sum_{m_j} p\left(m_j\right) \sum_{l_{m_j j}} p\left(s_j|l_{m_j j}\right) p\left(l_{m_j j}|c_{m_j j}\right) \\ &= \sum_{m_j} \frac{1}{N} \sum_{l_{m_j j}} \delta\left[l_{m_j j} = s_j\right] \frac{\delta\left[f_{m_j}\left(l_{m_j j}\right) = c_{m_j j}\right]}{\sum_{l=1}^{L} \delta\left[f_{m_j}(l) = c_{m_j j}\right]} \end{aligned} \tag{1}$$

$$= \frac{1}{N} \sum_{m_j=1}^{N} \frac{\delta\left[f_{m_j}\left(s_j\right) = c_{m_j j}\right]}{\sum_{l=1}^{L} \delta\left[f_{m_j}(l) = c_{m_j j}\right]}, \tag{2}$$

where $\delta[\cdot]$ is the Kronecker delta. The interpretation of Eq. (2) is simple: at each voxel j, every atlas equally spreads its vote over all the fine labels that are compatible with the coarse label c_{nj}. The segmentation is just the label that maximizes Eq. (2) with respect to s_j.

Table 1
Equations for the probabilistic model of generalized majority voting.

1. $\{\mathbf{l}_n\} \sim p(\{\mathbf{l}_n\}) \propto 1$
2. $\mathbf{m} \sim p(\mathbf{m}) \propto 1$
3. $c_{nj} = f_n(l_{nj}), \forall\, n, j$
4. $s_j = l_{m_j j}, \forall j$

Generalization of semi-locally weighted voting

Here we generalize the model from Sabuncu et al. (2010) to the scenario with collapsed labels. The graphical model is shown in Fig. 2, and the corresponding equations in Table 2. The model for the labels is the same as for majority voting; however, now there is a Markov random field (MRF) prior with a predefined constant β on the membership field $\mathbf{m}$, in order to encourage spatial patterns of labels observed in the training data. In addition to the segmentation, the membership field now also generates the intensities of the test scan at each voxel by selecting atlas m_j, taking its intensity $i_{m_j j}$, and corrupting it with Gaussian noise with variance σ^2. The variance is independent of the spatial location. In Sabuncu et al. (2010), it is set to a predefined value; here, we estimate it from the data instead assuming a flat prior for it. The segmentation is formulated as:

$$\begin{aligned} \hat{\mathbf{s}} &= \underset{\mathbf{s}}{\operatorname{argmax}}\, p(\mathbf{s}|\{\mathbf{i}_n\}, \{\mathbf{c}_n\}, \mathbf{y}) \\ &= \underset{\mathbf{s}}{\operatorname{argmax}} \int_{\sigma^2} p\left(\mathbf{s}\middle|\sigma^2, \{\mathbf{i}_n\}, \{\mathbf{c}_n\}, \mathbf{y}\right) p\left(\sigma^2|\{\mathbf{i}_n\}, \{\mathbf{c}_n\}, \mathbf{y}\right) d\sigma^2. \end{aligned} \tag{3}$$

$$\approx \underset{\mathbf{s}}{\operatorname{argmax}}\, p\left(\mathbf{s}|\hat{\sigma}^2, \{\mathbf{i}_n\}, \{\mathbf{c}_n\}, \mathbf{y}\right), \tag{4}$$

where we have used the mode approximation for the integral. The point estimate $\hat{\sigma}^2$ of the variance is:

$$\hat{\sigma}^2 = \underset{\sigma^2}{\operatorname{argmax}}\, p\left(\sigma^2|\{\mathbf{i}_n\}, \{\mathbf{c}_n\}, \mathbf{y}\right). \tag{5}$$

We now describe algorithms to compute the most likely variance $\hat{\sigma}^2 \hat{\sigma}^2$ 2.3.1 and subsequently the most likely segmentation.

Computation of the most likely variance $\hat{\sigma}^2$

We first note that, according to the model in Fig. 2, the variance σ^2 is independent of the coarse labels $\{\mathbf{c}_n\}$ when the segmentation

Table 2
Equations for the generative model of generalized semi-locally weighted voting. $\mathcal{H}_j$ represents the neighborhood of voxel j.

1. $\{\mathbf{l}_n\} \sim p(\{\mathbf{l}_n\}) \propto 1$
2. $\mathbf{m} \sim p(\mathbf{m}) \propto \prod_{j=1}^{J}, \exp\left[\frac{\beta}{2} \sum_{j' \in \mathcal{H}_j} \delta\left[m_j = m_{j'}\right]\right]$
3. $\sigma^2 \sim p(\sigma^2) \propto 1$
4. $c_{nj} = f_n(l_{nj}), \forall\, n, j$
5. $s_j = l_{m_j j}, \forall j$
6. $y_j \sim p\left(y_j|m_j, \sigma^2, i_{\cdot j}\right) = \mathcal{N}\left(y_j; i_{m_j j}, \sigma^2\right)$

s is unknown. Therefore, the problem in Eq. (5) can be rewritten as follows:

$$\begin{aligned}\hat{\sigma}^2 &= \underset{\sigma^2}{\operatorname{argmax}}\, p\left(\sigma^2|\{\mathbf{i}_n\},\{\mathbf{c}_n\},\mathbf{y}\right)\\ &= \underset{\sigma^2}{\operatorname{argmax}}\, p\left(\sigma^2|\{\mathbf{i}_n\},\mathbf{y}\right)\\ &= \underset{\sigma^2}{\operatorname{argmax}}\, \log p\left(\mathbf{y}|\sigma^2,\{\mathbf{i}_n\}\right).\end{aligned} \tag{6}$$

Eq. (6) requires marginalizing over **m**, which leads to an intractable sum due to the MRF prior. Instead, we will use the variational expectation maximization (VEM) algorithm to estimate an approximate solution. Rather than optimizing Eq. (6) directly, we maximize a lower bound J:

$$J\left(q\left(\mathbf{m},\sigma^2\right)\right) = \log p\left(\mathbf{y}\middle|\sigma^2,\{\mathbf{i}_n\}\right) - KL\left[q(\mathbf{m})\middle\|p\left(\mathbf{m}\middle|\mathbf{y},\sigma^2,\{\mathbf{i}_n\}\right)\right] \tag{7}$$

$$\begin{aligned}&= H[q] + \sum_{\mathbf{m}} q(\mathbf{m}) \log p\left(\mathbf{m},\mathbf{y}|\sigma^2,\{\mathbf{i}_n\}\right),\\ &\leq \log p\left(\mathbf{y}|\sigma^2,\{\mathbf{i}_n\}\right)\end{aligned} \tag{8}$$

where KL denotes the Kullback–Leibler divergence and H represents the entropy of a random variable. The inequality $J(q(\mathbf{m}, \sigma^2)) \leq \log p(\mathbf{y}|\sigma^2, \{\mathbf{i}_n\})$ holds because the KL divergence is non-negative. The distribution $q(\mathbf{m})$ represents an approximation to the posterior distribution of **m** given the observed intensities and the variance. This distribution is optimized over a class of restricted functions. The standard approximation, known as mean field approximation, is that q factorizes over voxels: $q(\mathbf{m}) = \prod_{j=1}^{J} q_j\left(m_j\right)$, where q_j is a categorical distribution over the indices of the atlases at voxel j.

VEM alternates between an expectation (E) step and a maximization (M) step. In the E step, we maximize the bound J with respect to $q(\mathbf{m})$. In the M step, we maximize J with respect to the model parameters – in this case, the variance σ^2. In the **E step**, it is convenient to work with Eq. (7): maximizing J is equivalent to minimizing the KL divergence, which yields the following update:

$$q_j\left(m_j\right) \propto p\left(y_j|i_{m_j j},\sigma^2\right) \exp\left[\beta \sum_{j'\in\mathcal{H}_j} q_{j'}\left(m_j\right)\right], \tag{9}$$

which can be solved with fixed point iterations, normalizing q_j after each step.

In the **M step**, it is more convenient to work with Eq. (8), since the entropy term can be disregarded. The maximization yields the following update:

$$\sigma^2 = \frac{1}{J}\sum_{j=1}^{J}\sum_{m_j=1}^{N} q_j\left(m_j\right)\left(y_j - i_{m_j j}\right)^2. \tag{10}$$

The VEM algorithm typically converges in a few (5–6) iterations. Note that, if we set $\beta = 0$ in the model, we recover the standard EM algorithm (Dempster et al., 1977).

Computation of the most likely segmentation $\hat{\mathbf{s}}$

Given $\hat{\sigma}^2$, computing the segmentation $\hat{\mathbf{s}}$ still requires evaluating an intractable sum over **m**. However, since $q(\mathbf{m})$ minimizes the KL divergence with $p(\mathbf{m}|\mathbf{y}, \sigma^2, \{\mathbf{i}_n\})$, we can approximate the problem in Eq. (4) with:

$$\begin{aligned}\hat{\mathbf{s}} &= \underset{\mathbf{s}}{\operatorname{argmax}} \sum_{\mathbf{m}} p(\mathbf{s}|\mathbf{m},\{\mathbf{c}_n\}) p\left(\mathbf{m}|\mathbf{y},\{\mathbf{i}_n\},\hat{\sigma}^2\right)\\ &\approx \underset{\mathbf{s}}{\operatorname{argmax}} \prod_{j=1}^{J}\sum_{m_j=1}^{N} q_j\left(m_j\right) p\left(s_j\middle|m_j,c_{\cdot j}\right).\end{aligned}$$

Therefore, the optimal segmentation can be computed voxel by voxel as:

$$\hat{s}_j \approx \underset{s_j}{\operatorname{argmax}} \sum_{m_j=1}^{N} q_j\left(m_j\right) p\left(s_j|c_{\cdot j},m_j\right),$$

which is almost identical to the right hand side of Eq. (1), with the difference that the term $p(m_j)$ has been replaced by $q_j(m_j)$. It is straightforward to show that the approximate posterior label probabilities are given by:

$$p\left(s_j|y_j,c_{\cdot j},i_{\cdot j},\hat{\sigma}^2\right) \approx \sum_{m_j=1}^{N} q_j\left(m_j\right) \frac{\delta\left[f_{m_j}\left(s_j\right) = c_{m_j j}\right]}{\sum_{l_{m_j j}} \delta\left[f_{m_j}\left(l_{m_j j}\right) = c_{m_j j}\right]}. \tag{11}$$

This expression is similar to the equation for the label posteriors of generalized majority voting (Eq. (2)). The difference is that the constant term $1/N$ has been replaced by the approximate membership posteriors q_j. This term depends on the image intensities, such that the contribution is higher for the atlases that are semi-locally more similar to the novel scan. The vote of each atlas is still spread equally among the fine labels they are compatible with the coarse label at each voxel.

Generalization of STAPLE

The generative model of STAPLE is as follows: the hidden segmentation is generated by a prior $p(\mathbf{s}) = \prod_{j=1}^{J} p(s_j)$, such that $p(s_j)$ is a categorical distribution that reflects the prior frequencies of the classes. In our scenario, we used a flat prior (i.e., $p(s_j) \propto 1$); preliminary experiments showed that the prior used in the original STAPLE algorithm – in which the frequencies are proportional to the volumes of the structures – considerably decreased the segmentation accuracy in small structures. Given the segmentation, the (deformed) atlas labels are assumed to be independent corrupted observations of the hidden ground truth segmentation. Each atlas has a corresponding confusion matrix through which its labels are generated, whereas the atlas intensities are disregarded in the fusion. The STAPLE algorithm first computes point estimates for the confusion matrices in light of the observed data, and then uses them to estimate the segmentation.

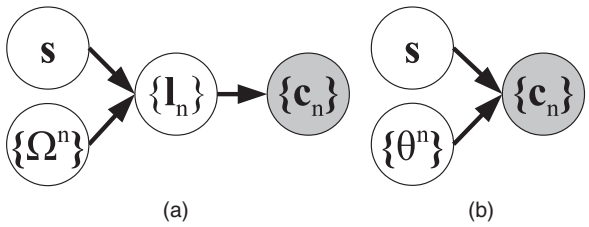

Fig. 3. Graphical models for generalized STAPLE. (a) Expanded version. (b) Compact version. Shaded variables are observed.

The graphical model for generalized STAPLE is shown in Fig. 3a. Each atlas is characterized by a confusion matrix Ω^n. The element Ω^n_{ls} corresponds to the probability that the true segmentation s is observed as label l by atlas n. The fine labels of the atlases are collapsed into the coarse, observed labels through the protocol functions f_n. For our purposes, it is actually more convenient to work with a more compact version of this model, as shown in Fig. 3b. We assume that the collapsed labels are generated directly by confusion matrices Θ^n: Θ^n_{cs} corresponds to the probability that the true segmentation s is observed as coarse label c by atlas n. The relation between Ω^n and Θ^n is simple: $\Theta^n_{cs} = \sum_{\{l|f_n(l)=c\}} \Omega^n_{ls}$. Note that the matrices $\{\Theta^n\}$ are not necessarily square, whereas the $\{\Omega^n\}$ are. Moreover, the matrices $\{\Theta^n\}$ will in general have different number of rows.

As in the original STAPLE algorithm, we compute the maximum likelihood estimates of the confusion matrices as follows:

$$\begin{aligned}\left\{\hat{\Theta}^n\right\} &= \underset{\{\Theta^n\}}{\operatorname{argmax}} \log p(\{\mathbf{c}_n\}|\{\Theta^n\}) \\ &= \underset{\{\Theta^n\}}{\operatorname{argmax}} \log \sum_{\mathbf{s}} p(\{\mathbf{c}_n\}|\mathbf{s},\{\Theta^n\}) \\ &= \underset{\{\Theta^n\}}{\operatorname{argmax}} \log \prod_{j=1}^{J} \sum_{s_j=1}^{L} \prod_{n=1}^{N} p\left(c_{nj}|s_j,\Theta^n\right) \\ &= \underset{\{\Theta^n\}}{\operatorname{argmax}} \sum_{j=1}^{J} \log \left[\sum_{s_j=1}^{L} \prod_{n=1}^{N} p\left(c_{nj}|s_j,\Theta^n\right)\right].\end{aligned} \tag{12}$$

Eq. (12) can be iteratively optimized with EM. In the **E step**, we compute $\hat{W}^s_j$, a soft classification of voxel j:

$$\widetilde{W}^s_j = \frac{\prod_{n=1}^{N} p\left(c_{nj}|s,\widetilde{\Theta}^n\right)}{\sum_{s'=1}^{L}\prod_{n=1}^{N} p\left(c_{nj}|s',\Theta^n\right)} = \frac{\prod_{n=1}^{N} \widetilde{\Theta}^n_{c_{nj}s}}{\sum_{s'=1}^{L}\prod_{n=1}^{N} \widetilde{\Theta}^n_{c_{nj}s'}}. \tag{13}$$

In the **M step**, we update the confusion matrices:

$$\Theta^n_{cs} = \frac{\sum_{j=1}^{J} \widetilde{W}^s_j \delta\left[c_{nj}=c\right]}{\sum_{j=1}^{J} \widetilde{W}^s_j}. \tag{14}$$

Once the EM algorithm has converged, the approximate label posteriors are given by:

$$p\left(s_j|c_{\cdot j}\right) \approx p\left(s_j|c_{\cdot j},\left\{\hat{\Theta}^n\right\}\right) = \widetilde{W}^{s_j}_j, \tag{15}$$

and the discrete segmentation is just:

$$\hat{s}_j \approx \underset{s}{\operatorname{argmax}}\, \widetilde{W}^s_j.$$

Shortcomings of the generalized methods

The presented generalizations of majority voting, semi-locally weighted voting and STAPLE suffer from several limitations. Generalized majority voting inherits from its parent the inability to exploit the information in the deformed atlas intensities. In addition, this method is unable to produce realistic label posteriors (soft segmentations) in many common multi-protocol scenarios due to the spreading of the votes across all the compatible fine labels at each location. For example, let's consider the problem of fusing the segmentation in Fig. 4a – which includes the cortex as a whole and most subcortical structures – with the segmentation in Fig. 4b – which includes the hippocampal subfields. Since the background label of the subfield data is compatible with all non-hippocampal labels, this scan would contribute a flat map of non-zero probability for all non-hippocampal structures all over the image domain. Even though the hard segmentation (Fig. 4c) might still be meaningful, the label posteriors given by the method (Fig. 4d and e) are not realistic. For instance, the posterior for the cortex is close to 0.5 (rather than 1) around this structure in Fig. 4d, and the same thing happens with the amygdala in Fig. 4e.

Soft segmentations are desirable for two reasons: first, they are useful to estimate the uncertainty in the boundary locations; and second, they enable us to compute more accurate estimates of the volumes of the different structures using expectations. If V_l is the volume of structure l in voxels (Iglesias et al., 2013):

$$E[V_l] = E\left[\sum_{j=1}^{J} \delta\left(s_j = l\right)\right] = \sum_{j=1}^{J} E\left[\delta\left(s_j = l\right)\right] = \sum_{j=1}^{J} p_j^{post}(l),$$

where p_j^{post} is the posterior label probability at voxel j, given by Eq. (2) (majority voting), 11 (semi-locally weighted voting) or 15 (STAPLE). If we computed volumes from the posteriors in Fig. 4d or e, the estimates would clearly not be accurate.

In contrast to generalized majority voting, generalized semi-locally weighted voting uses the image intensities to estimate the contributions from the different atlases; however, it still suffers from the shortcoming that it spreads the votes across all the fine labels and cannot generate meaningful posteriors. STAPLE, on the other hand, can produce realistic posteriors thanks to the nature of its generative model, but it shares with majority voting the limitation that it does not consider the image intensities in the fusion. Moreover, STAPLE was originally conceived as a method to merge manual segmentations, and generally performs poorly in the label fusion step of multi-atlas segmentation, where it is often outperformed even by majority voting (see for instance Iglesias et al. (2012)).

Even though some of these shortcomings could be addressed by newer versions of the algorithms (e.g., an extension of STAPLE using image intensities was presented in Asman and Landman (2013)), none of the generalized methods supports exchange of information between the atlases during the fusion. In the standard case where all

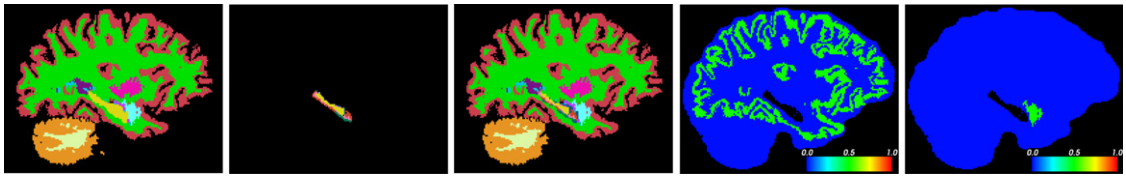

Fig. 4. Example to illustrate the shortcomings of generalized majority voting. (a) Labels for a registered atlas with labels for the whole cortex and for a number of subcortical structures. (b) Labels for a registered atlas with labels for the hippocampal subfields. (c) Fusion of (a) and (b) with generalized majority voting. (d) Posterior probability map for the cortex; note that the values are close to 0.5 (rather than 1) in the cortex, and that a large amount of probability mass is distributed all across the image in non-cortical areas (other than the hippocampus). (e) Posterior probability map for the amygdala; similar observations as for (d) can be made.

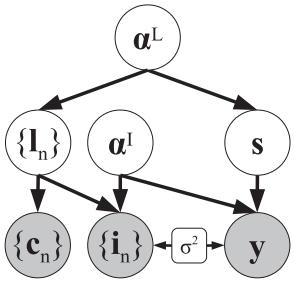

Fig. 5. Graphical models for MPLF, the proposed fusion method. Shaded variables are observed.

atlases have all possible labels, this is not necessary, since there is nothing they can learn from one another. However, in the multi-protocol scenario, the exchange of information between atlases with missing labels can improve the segmentation.

For instance, let us assume that we have three registered atlases $\{\mathbf{i}_1, \mathbf{c}_1\}$, $\{\mathbf{i}_2, \mathbf{c}_2\}$, $\{\mathbf{i}_3, \mathbf{c}_3\}$ and that, at a given voxel j, their deformed coarse labels are $c_{1j} = c'$, $c_{2j} = c''$ and $c_{3j} = c'''$, respectively. Let us further assume that c' and c'' correspond to fine labels l' and l'' without any ambiguity, i.e., $f_1(l) = c' \iff l = l'$ and $f_2(l) = c'' \iff l = l''$, but that c''' is the result of collapsing l' and l'', i.e., $f_3(l') = f_3(l'') = c'''$. In that case, if $i_{3j} \approx i_{1j}$, we would expect the hidden fine label l_{3j} to be l', whereas if $i_{3j} \approx i_{2j}$, we would expect it to be l'' instead.

To address the described issues, we introduce a label fusion method specifically designed for multi-protocol scenarios.

Proposed fusion method

Fig. 5 shows the graphical model of the proposed framework, which we coin "multi-protocol label fusion" (MPLF). Table 3 displays the corresponding equations. Essentially, we are assuming that both the registered atlases and the test scan are generated by a (latent) statistical atlas of labels ($\boldsymbol{\alpha}^{\mathcal{L}}$) and intensities ($\boldsymbol{\alpha}^{\mathcal{I}}$) in the space of the test scan. The assumption that the atlases were generated by the same statistical atlas is what allows the model to integrate the information across the atlases. Specifically, the vector $\boldsymbol{\alpha}_j^{\mathcal{L}} = \{\boldsymbol{\alpha}_{jl}^{\mathcal{L}}, l = 1, \ldots, L\}$ represents the a priori probability of observing the different labels (at the fine detail level) at voxel j. The vector $\boldsymbol{\alpha}_j^{\mathcal{I}} = \{\boldsymbol{\alpha}_{jl}^{\mathcal{I}}, l = 1, \ldots, L\}$ stores, for each fine label, the mean of a Gaussian that models the distribution of intensities at voxel j conditioned on that label. All the Gaussians share a predefined variance σ^2. The voxels are assumed to be independent of one another. As in the previous sections, the fine labels of the atlases are hidden, and we only have access to corresponding coarse segmentations $\{\mathbf{c}_n\}$.

Table 3
Equations for the generative model of Fig. 5. **Dir**[·] represents the Dirichlet distribution, and **1** is the all one vector.

1. $\boldsymbol{\alpha}^{\mathcal{L}} \sim p(\boldsymbol{\alpha}^{\mathcal{L}}) = \prod_{j=1}^{J} \mathrm{Dir}\left[\boldsymbol{\alpha}_j^{\mathcal{L}}; (1+\epsilon)\mathbf{1}\right]$
2. $l_{nj} \sim \boldsymbol{\alpha}_j^{\mathcal{L}}, \forall\, n, j$
3. $c_{nj} = f_n(l_{nj}), \forall\, n, j$
4. $s_j \sim \boldsymbol{\alpha}_j^{\mathcal{L}}, \forall\, j$
5. $\boldsymbol{\alpha}^{\mathcal{I}} \sim p(\boldsymbol{\alpha}^{\mathcal{I}}) = \prod_{j=1}^{J} \prod_{l=1}^{L} \mathcal{N}\left[\alpha_{jl}^{\mathcal{I}}; \mu_0, \frac{\sigma^2}{\epsilon}\right]$
6. $i_{nj} \sim \mathcal{N}\left(\boldsymbol{\alpha}_{jl_{nj}}^{\mathcal{I}}, \sigma^2\right), \forall n, j$
7. $y_j \sim \mathcal{N}\left(\alpha_{js}^{\mathcal{I}}, \sigma^2\right), \forall j$

The model is completed with priors for $\boldsymbol{\alpha}^{\mathcal{L}}$ and $\boldsymbol{\alpha}^{\mathcal{I}}$, which we assume factorize over voxels as well. We use conjugate priors: for $\boldsymbol{\alpha}_j^{\mathcal{L}}$, we assume a Dirichlet distribution with concentration parameter $1 + \epsilon$, i.e., we assume that we have ϵ prior observations for each class at each voxel. For $\boldsymbol{\alpha}_{jl}^{\mathcal{I}}$ we assume a Gaussian distribution with mean u_0 and variance σ^2/ϵ, which is equivalent to having ϵ priors observations with sample mean μ_0. In practice, ϵ is small and the only objective of these priors is to ensure the numerical stability of the algorithm.

Exact inference within this model would require marginalizing over the model parameters, i.e., the statistical atlas ($\boldsymbol{\alpha}^{\mathcal{L}}$, $\boldsymbol{\alpha}^{\mathcal{I}}$), which leads to an intractable integral. We use the assumption that the posterior distribution of the model parameters is heavily peaked to approximate:

$$p(\mathbf{s}|\mathbf{y}, \{\mathbf{i}_n\}, \{\mathbf{c}_n\}) \approx p\left(\mathbf{s}\middle|\hat{\boldsymbol{\alpha}}^{\mathcal{I}}, \hat{\boldsymbol{\alpha}}^{\mathcal{L}}, \mathbf{y}, \{\mathbf{i}_n\}, \{\mathbf{c}_n\}\right), \tag{16}$$

where the point estimates $\hat{\alpha}^{\mathcal{I}}$ and $\hat{\alpha}^{\mathcal{L}}$ are given by:

$$\left(\hat{\boldsymbol{\alpha}}^{\mathcal{L}}, \hat{\boldsymbol{\alpha}}^{\mathcal{I}}\right) = \underset{\boldsymbol{\alpha}^{\mathcal{L}}, \boldsymbol{\alpha}^{\mathcal{I}}}{\operatorname{argmax}}\, p\left(\boldsymbol{\alpha}^{\mathcal{L}}, \boldsymbol{\alpha}^{\mathcal{I}}|\mathbf{y}, \{\mathbf{i}_n\}, \{\mathbf{c}_n\}\right). \tag{17}$$

To compute the point estimates, it is convenient to notice that the test image can be considered an extra atlas, such that:

$$\begin{aligned} \mathbf{i}_{n+1} &= \mathbf{y}, \\ \mathbf{l}_{n+1} &= \mathbf{s}, \\ \mathbf{c}_{n+1} &= 1, \\ f_{n+1}(s) &= 1, \forall s, \end{aligned}$$

In other words, the test image is an additional atlas (index $N+1$) with a constant coarse label 1, which is compatible with all the labels at the fine level. Then, we can use Bayes's rule to rewrite the problem in Eq. (17) as:

$$\begin{aligned} \left(\hat{\boldsymbol{\alpha}}^{\mathcal{L}}, \hat{\boldsymbol{\alpha}}^{\mathcal{I}}\right) &= \underset{\boldsymbol{\alpha}^{\mathcal{L}}, \boldsymbol{\alpha}^{\mathcal{I}}}{\operatorname{argmax}}\, p\left(\{\mathbf{i}_n\}, \{\mathbf{c}_n\}\middle|\boldsymbol{\alpha}^{\mathcal{L}}, \boldsymbol{\alpha}^{\mathcal{I}}\right) p\left(\boldsymbol{\alpha}^{\mathcal{L}}, \boldsymbol{\alpha}^{\mathcal{I}}\right) \\ &= \underset{\boldsymbol{\alpha}^{\mathcal{L}}, \boldsymbol{\alpha}^{\mathcal{I}}}{\operatorname{argmax}}\, \log p\left(\boldsymbol{\alpha}^{\mathcal{L}}\right) + \log p\left(\boldsymbol{\alpha}^{\mathcal{I}}\right) + \sum_{n=1}^{N+1} \log p\left(\mathbf{i}_n, \mathbf{c}_n\middle|\boldsymbol{\alpha}^{\mathcal{L}}, \boldsymbol{\alpha}^{\mathcal{I}}\right) \\ &= \underset{\boldsymbol{\alpha}^{\mathcal{L}}, \boldsymbol{\alpha}^{\mathcal{I}}}{\operatorname{argmax}} \sum_{j=1}^{J} \log p\left(\boldsymbol{\alpha}_j^{\mathcal{L}}\right) + \sum_{j=1}^{J} \sum_{l=1}^{L} \log p\left(\alpha_{jl}^{\mathcal{I}}\right) + \ldots \\ &\quad \ldots + \sum_{n=1}^{N+1} \sum_{j=1}^{J} \log \sum_{l_{nj}=1}^{L} p\left(i_{nj}|\boldsymbol{\alpha}_j^{\mathcal{I}}, l_{nj}\right) p\left(c_{nj}|l_{nj}\right) p\left(l_{nj}|\boldsymbol{\alpha}_j^{\mathcal{L}}\right), \end{aligned} \tag{18}$$

where

$$\begin{aligned} p\left(\boldsymbol{\alpha}_j^{\mathcal{L}}\right) &= \frac{1}{Z_\epsilon} \prod_{l=1}^{L} \left(\alpha_{jl}^{\mathcal{L}}\right)^{\in}, \\ p\left(\alpha_{jl}^{\mathcal{I}}\right) &= \mathcal{N}\left(\alpha_{jl}^{\mathcal{I}}; \mu_0, \sigma^2/\in\right), \\ p\left(i_{nj}\middle|\boldsymbol{\alpha}_j^{\mathcal{I}}, l_{nj}\right) &= \mathcal{N}\left(i_{nj}; \alpha_{jl_{nj}}^{\mathcal{I}}, \sigma^2\right), \\ p\left(c_{nj}\middle|l_{nj}\right) &= \delta\left[f_n\left(l_{nj}\right) = c_{nj}\right], \\ p\left(l_{nj}\middle|\boldsymbol{\alpha}_j^{\mathcal{L}}\right) &= \alpha_{jl_{nj}}^{\mathcal{L}}. \end{aligned}$$

To solve the optimization problem of Eq. (18), we use an EM algorithm: we iteratively build a lower bound to the objective function of Eq. (18) that touches it at the current estimate of ($\boldsymbol{\alpha}^{\mathcal{L}}$, $\boldsymbol{\alpha}^{\mathcal{I}}$) (E step), and then optimize this bound with respect to ($\boldsymbol{\alpha}^{\mathcal{L}}$, $\boldsymbol{\alpha}^{\mathcal{I}}$) (M step).

In the **E step**, we make a soft assignment of each label l at the fine level of detail to each voxel j in each atlas n:

$$\widetilde{W}_{nj}^{l} = \frac{\mathcal{N}\left(i_{nj}; \boldsymbol{\alpha}_{jl}^{\mathcal{I}}, \sigma^2\right) \boldsymbol{\alpha}_{jl}^{\mathcal{L}}\, \delta\left[c_{nj} = f_n(l)\right]}{\sum_{l'} \mathcal{N}\left(i_{nj}; \alpha_{jl'}^{\mathcal{I}}, \sigma^2\right) \alpha_{jl'}^{\mathcal{L}}\, \delta\left[c_{nj} = f_n(l')\right]}. \tag{19}$$

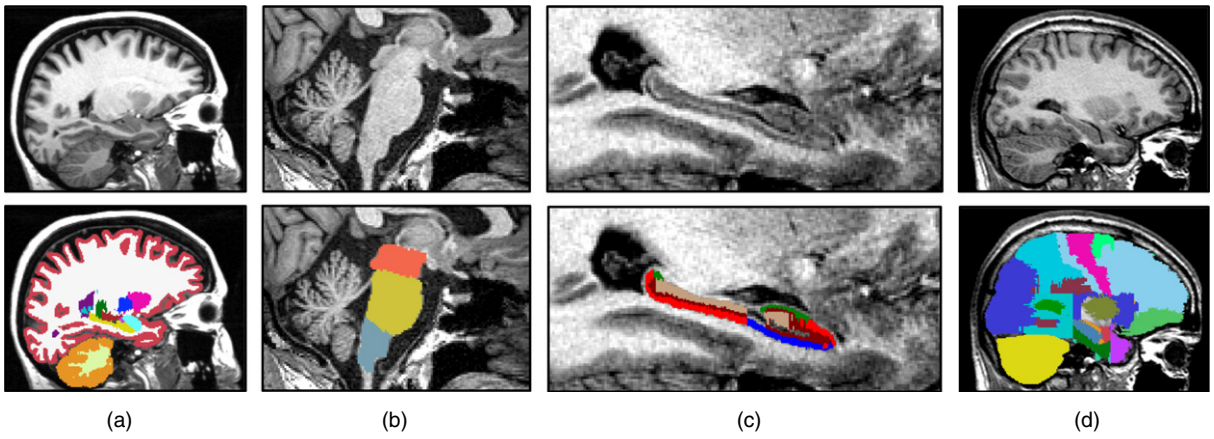

Fig. 6. Sample sagittal slices of the four datasets used in this study, with manual annotations overlaid: (a) FreeSurfer, (b) Brainstem, (c) Winterburn, (d) Hammers dataset. We only show a region of interest around the labels in (b,c) — the scans cover the whole brain.

Note that, if atlas n is labeled at the fine level already, then $(f_n)^{-1}(c_{nj})$ is unique and $\widetilde{W}^l_{nj} = \delta[(f_n)^{-1}(c_{nj}) = l]$. These soft assignments are used to form the lower bound:

$$Q\left(\boldsymbol{\alpha}^{\mathcal{L}}, \boldsymbol{\alpha}^{\mathcal{I}}; \widetilde{\boldsymbol{\alpha}}^{\mathcal{L}}, \widetilde{\boldsymbol{\alpha}}^{\mathcal{I}}\right) = -J \log Z_\epsilon - \sum_{n=1}^{N+1}\sum_{j=1}^{J}\sum_{l=1}^{L} \widetilde{W}^l_{nj} \log \widetilde{W}^l_{nj} + \ldots$$
$$\ldots + \sum_{j=1}^{J}\sum_{l=1}^{L}\left[\epsilon \log \alpha^{\mathcal{L}}_{jl} + \log \mathcal{N}\left(\alpha^{\mathcal{I}}_{jl}; \mu_0, \sigma^2/\epsilon\right)\right] + \ldots \quad (20)$$
$$\ldots + \sum_{n=1}^{N+1}\sum_{j=1}^{J}\sum_{l=1}^{L} \widetilde{W}^l_{nj} \log\left[\mathcal{N}\left(i_{nj}; \alpha^{\mathcal{I}}_{jl}, \sigma^2\right) \delta\left[f_n(l) = c_{nj}\right] \alpha^{\mathcal{L}}_{jl}\right],$$

where $\widetilde{\boldsymbol{\alpha}}^{\mathcal{L}}$ and $\widetilde{\boldsymbol{\alpha}}^{\mathcal{I}}$ are the current estimates of the parameters.

The **M step** updates can be derived as:

$$\alpha^{\mathcal{I}}_{jl} = \frac{\epsilon\mu_0 + \sum_{n=1}^{N+1} \widetilde{W}^l_{nj} i_{nj}}{\epsilon + \sum_{n'=1}^{N+1} \widetilde{W}^l_{n'j}}, \quad (21)$$

$$\alpha^{\mathcal{L}}_{jl} = \frac{\epsilon + \sum_{n=1}^{N+1} \widetilde{W}^l_{nj}}{\epsilon L + N + 1}. \quad (22)$$

Once the algorithm has converged, we can substitute the point estimates $\hat{\boldsymbol{\alpha}}^{\mathcal{L}}$ and $\hat{\boldsymbol{\alpha}}^{\mathcal{I}}$ back into Eq. (16) and use Bayes's rule to compute the segmentation. It is straightforward to show that the approximate posterior label probabilities of the voxels – which are independent of each other – are given by:

$$p\left(s_j | \hat{\boldsymbol{\alpha}}^{\mathcal{L}}_j, \hat{\boldsymbol{\alpha}}^{\mathcal{I}}_j, y_j, i_{\cdot j}, c_{\cdot j}\right) = \widetilde{W}^{s_j}_{N+1,j}. \quad (23)$$

Therefore, the optimal discrete segmentation is:

$$\hat{s}_j \approx \underset{s}{\operatorname{argmax}}\, \widetilde{W}^{s}_{N+1,j}, \quad (24)$$

and the expectation of the volume of label l is (in voxels):

$$E[V_l] = \sum_{j=1}^{J} \widetilde{W}^l_{N+1,j}. \quad (25)$$

Experiments and results

MRI data

We used four different datasets of manually labeled T1 MRI scans in this study (see sample slices in Fig. 6):

- FreeSurfer dataset: 39T1-weighted, 1 mm isotropic scans with 36 cortical and subcortical labels (see delineation protocol in Caviness et al. (1989)). The cerebral cortex and white matter are considered single entities. We note that these are the subjects that were used to train the probabilistic atlas in FreeSurfer (Fischl et al., 2002).
- Brainstem dataset: ten T1-weighted, 1 mm isotropic scans with manual labels for the medulla oblongata, pons and midbrain, i.e., the substructures of the brainstem. The delineation protocol is described in Iglesias et al. (submitted for publication).
- Winterburn dataset: five 0.6 mm isotropic scans[1] with annotations of the hippocampal subfields – subiculum, CA1, CA23, CA4 and molecular layer. The acquisition and manual delineation of the data are described in Winterburn et al. (2013). The dataset includes T1-weighted and T2-weighted scans; only the T1-weighted volumes were used here.
- Hammers dataset[2]: 20T1-weighted, 1 mm isotropic scans with 67 labels of cortical and subcortical structures. The labels of the cortical structures do not separate white from gray matter, and there is a single cerebellar label, which groups its gray and white matter. Further details on the dataset and on the manual labeling protocol can be found in Hammers et al. (2003); Gousias et al. (2008).

The scans from all four datasets have fields of view covering the whole brain. Additional details on the acquisition can be obtained from the corresponding publications.

Definition of fine labels and protocol functions

In the fusion, we defined a set of 148 labels given by all the possible intersections of regions defined in the four datasets; note that this

[1] In the original publication they work with 0.3 mm upsampled data, but the native resolution is 0.6 mm.

[2] www.brain-development.org. ©Copyright Imperial College of Science, Technology and Medicine 2007. All rights reserved.

process generates more regions than the total number of unique labels in the four datasets (102). Specifically, we defined two labels for each lobe in the Hammers dataset, one for the corresponding white matter and one for the corresponding cortex, based on their intersections with the gray and white matter in the FreeSurfer dataset. We also defined a number of labels to cope with differences in labeling protocols of the same structure; even though the protocols are in general very similar to each other, there are two exceptions. First, the Hammers hippocampus does not include the tail, while the FreeSurfer and Winterburn protocols do. To cope with this, we split up each of the five hippocampal subfields into an anterior (head/body) and a posterior (tail) region; this yields a total of 20 hippocampal labels – 10 per side. Second, the midbrain of the Brainstem dataset coincides with the superior part of the brainstem in the Hammers dataset, but extends further in the superior direction than the brainstem in FreeSurfer. To model this difference, we split the midbrain into an inferior and a superior part; the latter is further split into left and right. The midbrain label in the Brainstem dataset and the brainstem in the Hammers dataset include all three regions, while the FreeSurfer dataset considers the inferior region part of the brainstem and the superior regions part of the left and right diencephala, respectively.

Given these region definitions, we specified four unique protocol functions f (one per dataset, as illustrated in Fig. 7):

- FreeSurfer dataset: the protocol collapses all cortical structures into two generic cortex labels (left and right), all white matter structures (including the left and right corpus callosum, only defined in the Hammers dataset) into two white matter regions, all the hippocampal subfields into whole hippocampi, and the brainstem labels into a whole brainstem region – except for the superior midbrain regions, which are collapsed together with the diencephala.
- Brainstem dataset: the protocol collapses all the non-brainstem regions into a single background label.
- Winterburn dataset: the protocol collapses all the non-hippocampal regions into a generic background label.
- Hammers dataset: the protocol collapses the white and gray matter labels of each lobe into a single label, all the brainstem and hippocampal regions into two generic labels, and the white matter and cortex of the cerebellum into a single structure.

Experimental setup

All 74 brain scans were resampled to 1 mm isotropic resolution, skull stripped, bias field corrected and intensity normalized with FreeSurfer (Fischl et al., 2002). The intensity normalization is necessary because it enables us to directly compare image intensities (even across datasets), which is a critical assumption in semi-locally weighted fusion and MPLF. The scans were then pairwise registered with the software package ANTS (Avants et al., 2008). We used the default parameters for the deformation model (SyN[0.25]), number of iterations ($30 \times 90 \times 20$) and cost function (neighborhood cross correlation). However, we increased the radius of the window of the cost function from 4 to 6, and the regularization parameter from 3 to 4; these values better coped with the differences in intensity profile between datasets that cannot be completely eliminated by the intensity normalization step.

The four proposed label fusion schemes were used to segment the scans in a leave-one-out manner, where the test subject was left out from the atlas set $\{\mathbf{i}_n, \mathbf{l}_n\}$. The parameter setting and initialization were as follows. In semi-locally weighted fusion, we initialized $\sigma^2 = 100$, and set $\beta = 0.75$ – as in Sabuncu et al. (2010). In STAPLE, we initialized each column of each Θ^n such that 0.95 of the probability mass is equally distributed among the coarse labels that are compatible with the fine label corresponding to that column; the remaining 0.05 is equally spread across the non-compatible coarse labels. In MPLF, we set $\epsilon = 10^{-6}$, $\mu_0 = 65$ (which is the typical intensity of gray matter after normalization) and $\sigma^2 = 100$ (again, inspired by Sabuncu et al. (2010)).

We evaluated the label fusion methods in two different ways: directly with Dice scores, and indirectly through an aging experiment. First, we computed Dice scores between the manual delineations and the protocol-transformed, automated segmentations produced by the different label fusion methods, i.e., we compared $\mathbf{c}_{tj}$ with $f_t(\hat{\mathbf{s}}_j)$ (where we use the subscript t to refer to the test scan) for all four datasets. In addition, we also evaluated the performance of MPLF against the number of training scans. In this experiment, we used the FreeSurfer dataset, which has the most scans and therefore provides the largest range for the analysis. We computed segmentations of the FreeSurfer scans using N_{fs} other randomly selected FreeSurfer scans ("fs" stands for FreeSurfer) and a pool of (randomly selected) scans from the other datasets, such that the proportion of atlases from each of the datasets in the pool was approximately constant (see Table 4). This ensures that the performance depends on the number of atlases, rather than the proportion of scans from the FreeSurfer dataset in the pool of training images. The experiment was repeated 10 times for each N_{fs} with different random selections of scans except for the cases $N_{\mathrm{fs}} = 1$, where we used all 38 left-out scans in the dataset, and $N_{\mathrm{fs}} = 38$, where there is only one possible combination scans (since we are using all the available atlases).

In a second set of experiments, we evaluated MPLF indirectly by analyzing the effect of age on the median thickness of each cortical region, as well as on the volume of the different subcortical structures. This analysis was carried at the fine level of label detail, so we measured volumes and thicknesses of structures that were not defined in any of the training datasets. Both the thicknesses and the volumes were computed from the soft segmentations: the volumes were computed with Eq. (25), and the thicknesses were estimated from the label posteriors (Eq. (23)) using the algorithm described in (Aganj et al., 2009). This technique estimates the thickness at a point of interest by minimizing the line integral over the probability map of the gray matter on line segments centered at that point. For each cortical region, we took the median (a robust estimate) of the thicknesses given by this method for the voxels belonging to that region, as estimated by the hard

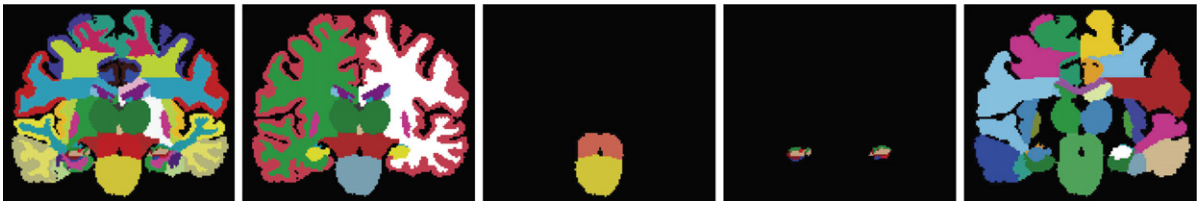

Fig. 7. Protocol functions: (a) sagittal slice of a segmentation at the fine label level; (b) effect of applying protocol function corresponding to FreeSurfer dataset; (c) Brainstem dataset; (d) Winterburn dataset; (e) Hammers dataset. It is the goal of this study to produce segmentations such as (a) by merging segmentations like (b–e).

Table 4
Number of atlases from each dataset in the training pool for the experiment testing the performance of MPLF against the number of training scans.

FreeSurfer (N_{fs})	1	3	5	8	16	23	31	37
Brainstem	1	1	2	2	4	6	8	10
Winterburn	1	1	1	1	2	3	4	5
Hammers	1	2	3	4	8	12	16	20

segmentation from Eq. (24). Note that both the estimation of the volume and of the cortical thickness from soft segmentations require faithful posteriors; probability maps like the ones showed in Figs. 4c or c would yield unrealistic estimates.

The aging experiment was performed on the FreeSurfer dataset, which has the most subjects and the widest age range (53.3 ± 23.3 years). For each subcortical brain structure, we first fitted a generalized linear model (GLM) predicting its volume as a linear combination of the age of the subject, his intracranial volume (as estimated by FreeSurfer) and a bias. Then, for each cortical structure, we fitted a GLM predicting its median thickness as a linear combination of the age of the subject and a bias. Finally, a statistical *t*-test was used to assess whether the coefficient related to age in each GLM was significantly different from zero. In order to increase the power of the analysis in the subcortical structures, we left-right averaged their volumes – for the median cortical thickness this is not as advantageous, since it is a robust estimate already.

Results

Direct validation: Dice scores

Table 5 shows the mean Dice scores across the structures defined within each dataset. Generalized STAPLE produces very variable results, yielding excellent segmentation for some structures but poor outputs for others (e.g., brainstem and Winterburn datasets). On average, it outperforms generalized majority voting by 2% Dice. Generalized semi-locally weighted voting takes advantage of the image intensities of the deformed atlases to yield an average Dice 1% higher than that of generalized STAPLE. MPFL clearly outperforms all the other methods by communicating information between the atlases: its average Dice is 3% higher than that of the second best method (generalized semi-locally weighted voting). Fig. 8 shows the mean Dice score produced by MPFL in the FreeSurfer dataset as a function of the number of training scans. The plot shows that MPFL only requires 3 atlases to yield Dice scores similar to those of generalized semi-locally weighted voting with 38 atlases. The performance of MPFL saturates at approximately 30 scans.

Fig. 9 shows box plots for the Dice scores between the manual and automated segmentations for each structure in the four datasets. In the FreeSurfer dataset, generalized majority voting performs satisfactorily for most structures except for the cortex, which is difficult to register. Semi-local weighting provides a boost in the performance for some of the structures, particularly the caudate and the cortex, thanks to the use of image intensity information in the fusion. Generalized STAPLE produces very variable results: for some structures it outperforms semi-locally weighted fusion (diencephalon, amygdala, pallidum, putamen), but for others it produces rather poor results (most notably the cortex). MPLF successfully combines all the training data to produce the highest Dice scores for every structure other than the cerebellum.

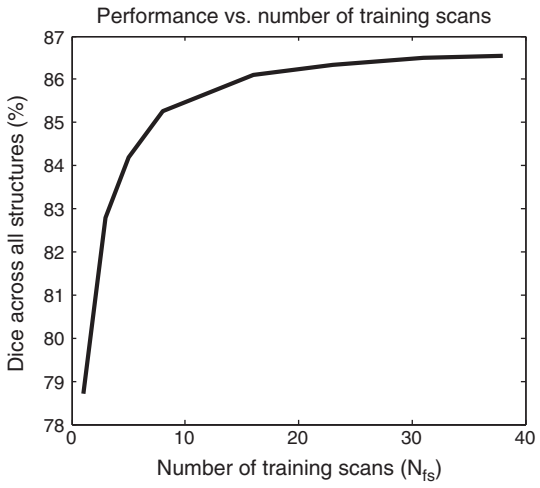

Fig. 8. Mean Dice score (in %) for the FreeSurfer dataset as a function of the number of training scans of the FreeSurfer dataset $\boldsymbol{N_{fs}}$. The pool of training scans also includes atlases from the other three datasets, such that the ratio of scans from the different datasets is approximately constant (see Table 4). The Dice scores are averaged over all structures and ten random selections of scans (except for $\boldsymbol{N_{fs}} = 1$, where we used all 38 left-out scans, and $\boldsymbol{N_{fs}} = 38$, where there is only one possible combination of atlases).

In absolute terms, the results from MPLF in subcortical structures are slightly worse than those reported in the literature by state-of-the-art fusion algorithms working in single-protocol settings (e.g., Sabuncu et al., 2010). This is not caused by shortcomings of our algorithm, but rather by the fact that it produces a segmentation that "averages" subtle differences in labeling protocols that are not explicitly modeled in our fusion framework. Therefore, the automatic segmentation does not necessarily agree perfectly with any of the individual protocols, and is penalized when computing Dice scores against the ground truth segmentations of the individual datasets. For example, this is the reason – in addition to differences in registrations – for the discrepancy between our results for generalized majority voting in the Hammers dataset and the results reported in Heckemann et al. (2006) using the same brain scans.

In cortical structures, MPLF shares the limitation of other label fusion algorithms (and registration based segmentation methods in general), that volumetric registration of the cerebral cortex is extremely difficult. The cortical segmentation could possibly be improved by replacing ANTs with a registration algorithm specifically designed for the cortex, e.g., (Postelnicu et al., 2009).

The same trends that were observed in the FreeSurfer dataset apply to the Brainstem, Winterburn and Hammers datasets. Semi-locally weighted voting provides a small improvement over majority voting, STAPLE yields very variable results (rather poor, in some cases) and MPLF outperforms the other three for all the structures of interest, other than CA1 in the hippocampus and the cerebellum (and some lobes/gyri) in the Hammers dataset. In absolute terms, the Dice scores are high in the Brainstem dataset – though difficult to place in a global context due to the lack of midbrain, pons and medulla segmentation algorithms in the literature. On the other hand, they are low for the hippocampal subfields, due to their thin shapes (all but CA4) and the insufficient resolution to segment them accurately. In the Hammers dataset, Dice scores of the subcortical structures are one notch below the results

Table 5
Mean Dice score (in %) across all structures within each dataset, as well as for all structures from all datasets combined. "Maj. vot." represents generalized majority voting, "SL weight." represents generalized semi-locally weighted fusion, and "G-STAPLE" represents generalized STAPLE.

Dataset	Maj. vot.	SL weight.	G-STAPLE	MPLF
FreeSurfer	80.6	82.6	82.6	86.5
Brainstem	85.2	85.6	74.2	86.9
Winterburn	43.1	46.8	36.4	50.5
Hammers	75.1	78.6	78.2	81.6
Combined	77.0	79.8	79.1	83.1

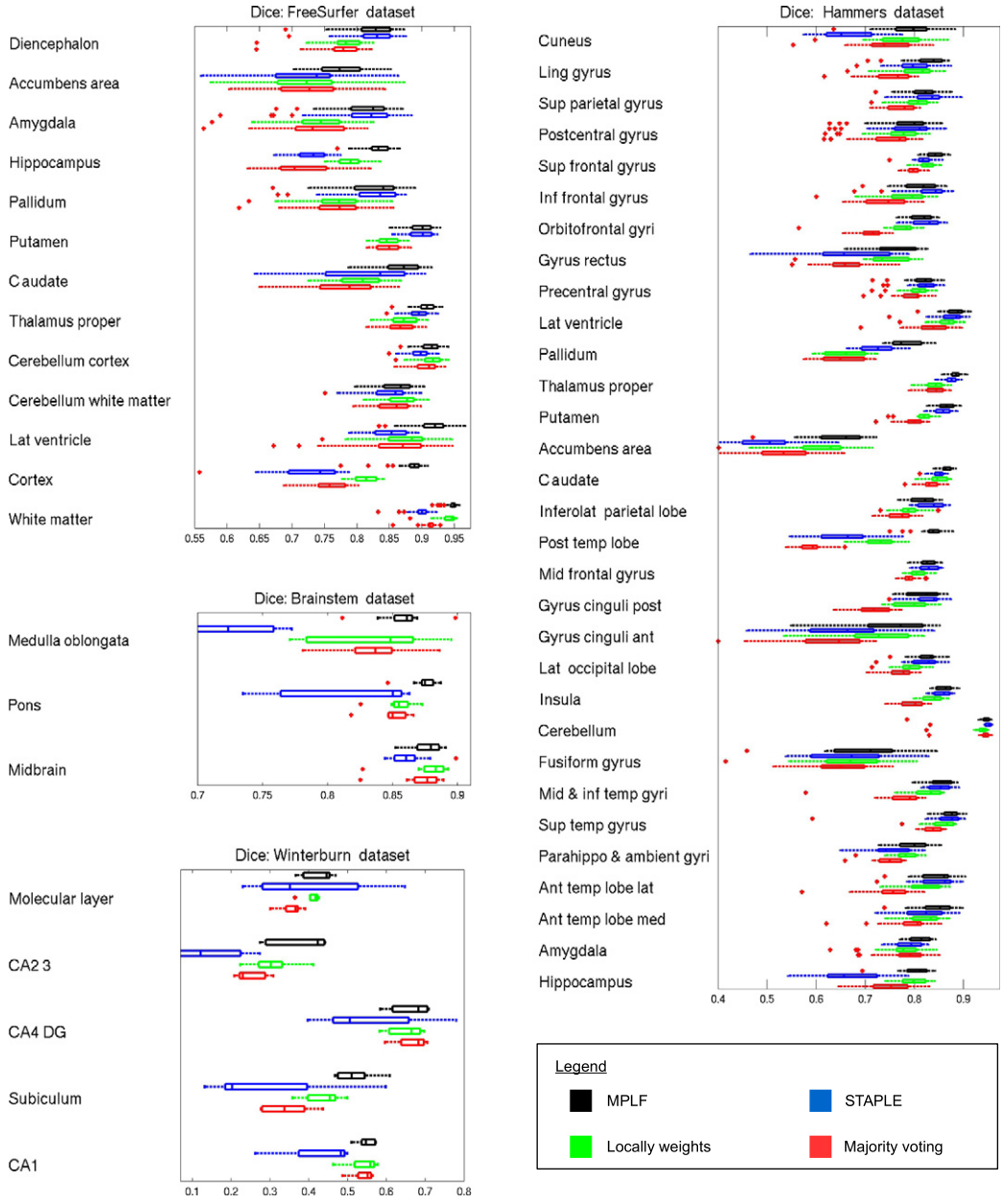

Fig. 9. Box plots for the Dice scores corresponding to the four datasets. The central mark is the median, the box spans from the first to the third quartile, and the whiskers span the extreme data points not considered outliers (which are marked with red crosses).

from the FreeSurfer dataset; this is possibly due to the fact that the averaging in labeling protocols is skewed towards the FreeSurfer data due to the larger presence in the training dataset – twice as many.

Fig. 10 shows segmentations and 3D renderings of a sample test scan from the FreeSurfer dataset. The automated segmentations in Fig. 10c–f are much richer than the manual labels in Fig. 10b. For instance, the parcellated pial and white matter surfaces (Fig. 10g–h) could not have been generated using any of the training datasets independently. Since they do not consider image intensities in the fusion, the generalizations of majority voting and STAPLE do not correctly segment the cortex, which is difficult to register. They also oversegment structures such as the pallidum. Generalized semi-locally weighted fusion ameliorates these problems through the use of intensities – particularly the cortical segmentation. However, it is outperformed by MPLF, which produces more accurate segmentations for the thalamus, pallidum, putamen and choroid plexus – while providing meaningful estimates of the label posteriors.

Indirect validation: aging study

The association between morphometric measurements and age (Table 6) shows strong consistency with prior work using other procedures. For example, we found strong associations between age and cortical thickness in frontal lobe regions including the superior frontal and precentral gyrus, weaker association with medial temporal regions, and strong reductions in volume of the thalamus with more moderate reductions in volume of the hippocampus (Walhovd et al., 2005; Salat et al., 2004). Even though the resolution of the scans was not sufficient to clearly distinguish among hippocampal subfields, we observed a larger effect of aging on CA1 and CA4-DG, consistent with prior work (Mueller and Weiner, 2009). The results on the brainstem also showed

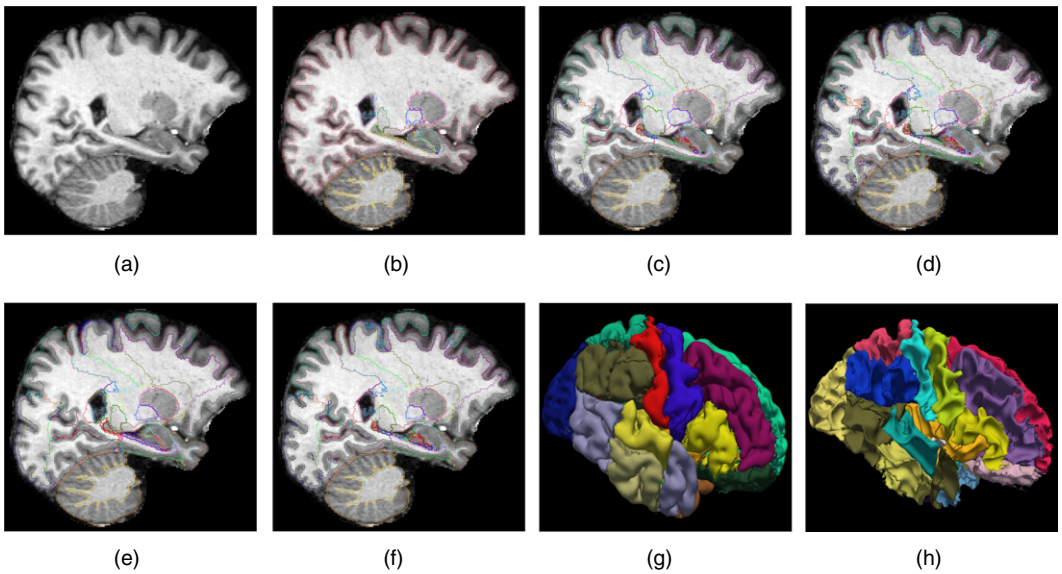

Fig. 10. (a) Sagittal slice of a scan from the FreeSurfer dataset. (b) Corresponding manual segmentation. (c) Automated segmentation with generalized majority voting. (d) Generalized semi-locally weighted segmentation. (e) Generalized STAPLE. (f) MPLF. (g) 3D rendering of the pial surface using MPLF. (h) Rendering of the white matter surface.

strong agreement with prior studies: the volume of the midbrain steadily decreases with age, the pons is spared, and the medulla suffers from minimal atrophy (Raininko et al., 1994).

Conclusion and discussion

In this paper, we have generalized three popular label fusion methods to scenarios where the atlases have been manually traced with different protocols. We have discussed the limitations of the generalized methods and proposed MPLF, an alternative algorithm that was shown to outperform them: the average Dice score on four datasets improved between 3% and 6% with respect to the generalizations. The core of both the generalized methods and MPLF is the definition of protocol functions that group sets of fine labels (hidden) into coarser labels (observed). Adding the coarse labels to the corresponding graphical models, the generalization of the three existing methods to the multi-protocol scenario is straightforward.

The extensions of majority voting, semi-locally weighted fusion and STAPLE are easy to implement and require very few changes with respect to the original algorithms. On the other hand, MPLF is computationally more expensive, since it requires optimizing a function with the EM algorithm at each voxel. In any case, the computational cost of the fusion (approximately one hour in our experiments) is small compared with the cost of nonlinearly registering the atlases (approximately one and a half hours per atlas, tens of hours in total), and the algorithm can be easily parallelized if necessary – since the voxels of the input scan are processed independently.

Future work will follow three directions. First, we will generalize newer, more sophisticated label fusion algorithms and compare them with MPLF. In particular, it will be interesting to consider extensions of STAPLE that support soft labels and spatially varying performance parameters. Second, we will consider the possibility of placing smoothness priors on the intensities and label probabilities of the statistical atlas in MPLF, as well as on the segmentation of the test scan. Even though the automated segmentations were accurate and smooth in our experiments, smoothness constraints might be important when the number of atlases is not as high as in this study. And third, we will generalize MPLF to cross-modality scenarios, which will introduce the capability to handle microscopic images (e.g., BigBrain[3] (Amunts et al., 2013)) or optical coherence tomography (Magnain et al., 2014), in order to model with very fine detail brain areas that are not visible with MRI.

As the amount of publicly available, heterogeneously labeled data continues to grow, we believe that segmentation methods that can cope with different protocols – such as the one we have described – will become increasingly important.

Acknowledgments

Support for this research was provided in part by the National Center for Research Resources (U24 RR021382, P41-RR14075, 1KL2RR025757-01), the National Institute for Biomedical Imaging and Bioengineering (P41-EB015896, R01EB006758, R01EB013565, 1K25EB013649-01), the National Institute on Aging (AG022381, 5R01AG008122-22, R01 AG016495-11), the National Center for Alternative Medicine (RC1 AT005728-01), the National Institute for Neurological Disorders and Stroke (R01 NS052585-01, 1R21NS 072652-01, 1R01NS070963, R01NS083534), and was made possible by the resources provided by Shared Instrumentation Grants 1S10RR023401, 1S10RR019307, and 1S10RR023043. Additional support was provided by The Autism & Dyslexia Project funded by the Ellison Medical Foundation, and by the NIH Blueprint for Neuroscience Research (5U01-MH093765), part of the multi-institutional Human Connectome Project. This research was also funded by TEKES (ComBrain), Harvard Catalyst, and financial contributions from Harvard and affiliations. JEI was supported by the Gipuzkoako Foru Aldundia (Fellows Gipuzkoa Program). MRS was partially supported by a BrightFocus grant (AHAF-A2012333). In addition, BF has a financial interest in CorticoMetrics, a company whose medical pursuits focus on brain imaging and measurement technologies. BF's interests were reviewed and are managed by Massachusetts General Hospital and Partners HealthCare in accordance with their conflict of interest policies.

[3] https://bigbrain.loris.ca.

Table 6
Effect of age: p-values (given as $-\log_{10} \boldsymbol{p}$) for *t*-test on significance of slope in the generalized linear model. (a) Volume of subcortical structures (left-right averaged). (b) Cortical thickness. Significant p values (below 0.05 after Bonferroni correction) are in bold. The threshold for significance at $\boldsymbol{\alpha} = 0.05$ is, after Bonferroni correction, $-\log_{10} \boldsymbol{p} \approx 2.65$.

(a)

Structure	$-\log_{10} p$
Lat. ventricle	**7.51**
Cerebellum WM	0.43
Cerebellum CT	2.37
Thalamus	**13.70**
Caudate	0.72
Putamen	**6.8**
Pallidum	**6.56**
Whole hippo.	**3.72**
Amygdala	**3.98**
Accumbens	**3.99**
Diencephalon	**10.27**
3rd ventricle	**7.10**
4th ventricle	0.59
Corp. call.	0.29
Whole brainst.	0.70
Cerebral WM	1.13
CA1	**2.84**
CA23	2.42
CA4-DG	**3.45**
Subiculum	1.38
Molec. layer	2.57
Midbrain	**6.14**
Pons	0.25
Medulla obl.	**2.68**

(b)

Cortical region	$-\log_{10} p$ (left side)	$-\log_{10} p$ (right side)
Ant. temp. lobe medial	**4.14**	**4.97**
Ant. temp. lobe lateral	**8.30**	**7.91**
Parahip. & ambient gyri	0.60	1.66
Sup. temp. gyrus	**9.26**	**11.21**
Middle & inf. temp. gyri	**4.15**	**4.86**
Fusiform gyrus	1.85	**2.94**
Insula	**8.90**	**9.49**
Lat. occipital lobe	**4.41**	**4.92**
Gyrus cinguli ant. part	**7.04**	**8.10**
Gyrus cinguli post. part	**9.10**	**9.84**
Middle front. gyrus	**11.69**	**10.94**
Post. temp. lobe	**7.25**	**8.69**
Inferolat. pariet. lobe	**12.71**	**10.61**
Precentral gyrus	**11.30**	**10.37**
Gyrus rectus	1.367	0.84
Orbitofrontal gyri	**8.82**	**11.93**
Inf. front. gyrus	**11.73**	**12.73**
Sup. front. gyrus	**11.87**	**13.03**
Postcentral gyrus	**8.44**	**8.35**
Sup. parietal gyrus	**6.63**	**8.20**
Lingual gyrus	**6.10**	**6.34**
Cuneus	**5.33**	**7.54**

References

Aganj, I., Sapiro, G., Parikshak, N., Madsen, S.K., Thompson, P.M., 2009. Measurement of cortical thickness from MRI by minimum line integrals on soft-classified tissue. Hum. Brain Mapp. 30 (10), 3188–3199.

Akhondi-Asl, A., Warfield, S., 2013. Simultaneous truth and performance level estimation through fusion of probabilistic segmentations. IEEE Trans. Med. Imaging 32 (10), 1840.

Aljabar, P., Heckemann, R.A., Hammers, A., Hajnal, J.V., Rueckert, D., 2009. Multi-atlas based segmentation of brain images: atlas selection and its effect on accuracy. NeuroImage 46 (3), 726–738.

Amunts, K., Lepage, C., Borgeat, L., Mohlberg, H., Dickscheid, T., Rousseau, M.-É., Bludau, S., Bazin, P.-L., Lewis, L.B., Oros-Peusquens, A.-M., et al., 2013. Bigbrain: an ultrahigh-resolution 3D human brain model. Science 340 (6139), 1472–1475.

Artaechevarria, X., Muñoz-Barrutia, A., Ortiz-de Solorzano, C., 2008. Efficient classifier generation and weighted voting for atlas-based segmentation: two small steps faster and closer to the combination oracle. Medical Imaging, International Society for Optics and Photonics (69141W–69141W).

Asman, A.J., Landman, B.A., 2012. Formulating spatially varying performance in the statistical fusion framework. IEEE Trans. Med. Imaging 31 (6), 1326–1336.

Asman, A.J., Landman, B.A., 2013. Non-local statistical label fusion for multi-atlas segmentation. Med. Image Anal. 17 (2), 194–208.

Avants, B.B., Epstein, C.L., Grossman, M., Gee, J.C., 2008. Symmetric diffeomorphic image registration with cross-correlation: evaluating automated labeling of elderly and neurodegenerative brain. Med. Image Anal. 12 (1), 26–41.

Awate, S., Whitaker, R., 2014. Multiatlas segmentation as nonparametric regression. IEEE Trans. Med. Imaging 33 (9), 1803–1817.

Cardoso, M.J., Leung, K., Modat, M., Keihaninejad, S., Cash, D., Barnes, J., Fox, N.C., Ourselin, S., 2013. Steps: similarity and truth estimation for propagated segmentations and its application to hippocampal segmentation and brain parcelation. Med. Image Anal. 17 (6), 671–684.

Caviness Jr., V., Filipek, P., Kennedy, D., 1989. Magnetic resonance technology in human brain science: blueprint for a program based upon morphometry. Brain Dev. 11 (1), 1–13.

Cline, H.E., Lorensen, W.E., Kikinis, R., Jolesz, F., 1990. Three-dimensional segmentation of MR images of the head using probability and connectivity. J. Comput. Assist. Tomogr. 14 (6), 1037–1045.

Commowick, O., Akhondi-Asl, A., Warfield, S.K., 2012. Estimating a reference standard segmentation with spatially varying performance parameters: local map staple. IEEE Trans. Med. Imaging 31 (8), 1593–1606.

Coupé, P., Manjón, J.V., Fonov, V., Pruessner, J., Robles, M., Collins, D.L., 2011. Patch-based segmentation using expert priors: application to hippocampus and ventricle segmentation. NeuroImage 54 (2), 940–954.

De Jong, L., Van der Hiele, K., Veer, I., Houwing, J., Westendorp, R., Bollen, E., De Bruin, P., Middelkoop, H., Van Buchem, M., Van Der Grond, J., 2008. Strongly reduced volumes of putamen and thalamus in Alzheimer's disease: an MRI study. Brain 131 (12), 3277–3285.

Dempster, A.P., Laird, N.M., Rubin, D.B., et al., 1977. Maximum likelihood from incomplete data via the EM algorithm. J. R. Stat. Soc. 39 (1), 1–38.

Duc, A.K.H., Modat, M., Leung, K.K., Cardoso, M.J., Barnes, J., Kadir, T., Ourselin, S., Initiative, A.D.N., et al., 2013. Using manifold learning for atlas selection in multi-atlas segmentation. PLoS One 8 (8), e70059.

Fischl, B., Salat, D., Busa, E., Albert, M., Dieterich, M., Haselgrove, C., van der Kouwe, A., Killiany, R., Kennedy, D., Klaveness, S., Montillo, A., Makris, N., Rosen, B., Dale, A., 2002. Whole brain segmentation: automated labeling of neuroanatomical structures in the human brain. Neuron 33, 341–355.

Gousias, I.S., Rueckert, D., Heckemann, R.A., Dyet, L.E., Boardman, J.P., Edwards, A.D., Hammers, A., 2008. Automatic segmentation of brain MRIs of 2-year-olds into 83 regions of interest. NeuroImage 40 (2), 672–684.

Hammers, A., Allom, R., Koepp, M.J., Free, S.L., Myers, R., Lemieux, L., Mitchell, T.N., Brooks, D.J., Duncan, J.S., 2003. Three-dimensional maximum probability atlas of the human brain, with particular reference to the temporal lobe. Hum. Brain Mapp. 19 (4), 224–247.

Heckemann, R., Hajnal, J., Aljabar, P., Rueckert, D., Hammers, A., 2006. Automatic anatomical brain MRI segmentation combining label propagation and decision fusion. NeuroImage 33 (1), 115–126.

Iglesias, J., Sabuncu, M., Van Leemput, K., 2012. A generative model for multi-atlas segmentation across modalities. 9th IEEE International Symposium on Biomedical Imaging (ISBI), pp. 888–891.

Iglesias, J.E., Sabuncu, M.R., Van Leemput, K., 2013. Improved inference in Bayesian segmentation using Monte Carlo sampling: application to hippocampal subfield volumetry. Med. Image Anal. 17 (7), 766–778.

Iglesias, J., Van Leemput, K., Bhatt, P., Casillas, C., Dutt, S., Boxer, A., Fischl, B., 2014. Bayesian Segmentation of Brainstem Structures in MRI (manuscript submitted for publication).

Isgum, I., Staring, M., Rutten, A., Prokop, M., Viergever, M., van Ginneken, B., 2009. Multi-atlas-based segmentation with local decision fusion — application to cardiac and aortic segmentation in CT scans. IEEE Trans. Med. Imaging 28 (7), 1000–1010.

Klein, S., Staring, M., Murphy, K., Viergever, M.A., Pluim, J.P., 2010. Elastix: a toolbox for intensity-based medical image registration. IEEE Trans. Med. Imaging 29 (1), 196–205.

Knickmeyer, R.C., Gouttard, S., Kang, C., Evans, D., Wilber, K., Smith, J.K., Hamer, R.M., Lin, W., Gerig, G., Gilmore, J.H., 2008. A structural MRI study of human brain development from birth to 2 years. J. Neurosci. 28 (47), 12176–12182.

Landman, B.A., Bogovic, J.A.., Prince, J.L., 2009. Efficient anatomical labeling by statistical recombination of partially label datasets. Proc. of ISMRM, p. 269.

Landman, B.A., Bogovic, J.A., Prince, J.L., 2010. Simultaneous truth and performance level estimation with incomplete, over-complete, and ancillary data. SPIE Medical Imaging, International Society for Optics and Photonics (76231N–76231N).

Landman, B.A., Asman, A.J., Scoggins, A.G., Bogovic, J.A., Xing, F., Prince, J.L., 2012. Robust statistical fusion of image labels. IEEE Trans. Med. Imaging 31 (2), 512–522.

Magnain, C., Augustinack, J.C., Reuter, M., Wachinger, C., Frosch, M.P., Ragan, T., Akkin, T., Wedeen, V.J., Boas, D.A., Fischl, B., 2014. Blockface histology with optical coherence tomography: a comparison with Nissl staining. NeuroImage 84, 524–533.

Mueller, S.G., Weiner, M.W., 2009. Selective effect of age, apo e4, and Alzheimer's disease on hippocampal subfields. Hippocampus 19 (6), 558–564.

Postelnicu, G., Zollei, L., Fischl, B., 2009. Combined volumetric and surface registration. IEEE Trans. Med. Imaging 28 (4), 508–522.

Raininko, R., Autti, T., Vanhanen, S.-L., Ylikoski, A., Erkinjuntti, T., Santavuori, P., 1994. The normal brain stem from infancy to old age. Neuroradiology 36 (5), 364–368.

Rohlfing, T., Brandt, R., Menzel, R., Maurer Jr., C.R., 2004. Evaluation of atlas selection strategies for atlas-based image segmentation with application to confocal microscopy images of bee brains. NeuroImage 21 (4), 1428–1442.

Sabuncu, M., Yeo, B., Van Leemput, K., Fischl, B., Golland, P., 2010. A generative model for image segmentation based on label fusion. IEEE Trans. Med. Imaging 29, 1714–1729.
Salat, D.H., Buckner, R.L., Snyder, A.Z., Greve, D.N., Desikan, R.S., Busa, E., Morris, J.C., Dale, A.M., Fischl, B., 2004. Thinning of the cerebral cortex in aging. Cereb. Cortex 14 (7), 721–730.
Walhovd, K.B., Fjell, A.M., Reinvang, I., Lundervold, A., Dale, A.M., Eilertsen, D.E., Quinn, B.T., Salat, D., Makris, N., Fischl, B., 2005. Effects of age on volumes of cortex, white matter and subcortical structures. Neurobiol. Aging 26 (9), 1261–1270.
Wang, H., Suh, J.W., Das, S.R., Pluta, J.B., Craige, C., Yushkevich, P.A., 2013. Multi-atlas segmentation with joint label fusion. IEEE Trans. Pattern Anal. Mach. Intell. 35 (3), 611–623.
Warfield, S.K., Zou, K.H., Wells, W.M., 2004. Simultaneous truth and performance level estimation (staple): an algorithm for the validation of image segmentation. IEEE Trans. Med. Imaging 23 (7), 903–921.
Winterburn, J.L., Pruessner, J.C., Chavez, S., Schira, M.M., Lobaugh, N.J., Voineskos, A.N., Chakravarty, M.M., 2013. A novel in vivo atlas of human hippocampal subfields using high-resolution 3T magnetic resonance imaging. NeuroImage 74, 254–265.
Yendiki, A., Panneck, P., Srinivasan, P., Stevens, A., Zöllei, L., Augustinack, J., Wang, R., Salat, D., Ehrlich, S., Behrens, T., Jbabdi, S., Gollub, R., Fischl, B., 2011. Automated probabilistic reconstruction of white-matter pathways in health and disease using an atlas of the underlying anatomy. Front. Neuroinformatics 5 (23), 1–12.